CROSS DOMAIN ENSEMBLE DISTILLATION FOR DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

For domain generalization, the task of learning a model that generalizes to unseen target domains utilizing multiple source domains, many approaches explicitly align the distribution of the domains. However, the optimization for domain alignment has a risk of overfitting since the target domain is not available. To address the issue, this paper proposes a method for domain generalization by employing self-distillation. The proposed method aims to train a model robust to domain shift by allowing meaningful erroneous predictions in multiple domains. Specifically, our method matches the ensemble of predictive distributions of data with the same class label but different domains with each predictive distribution. We also propose a de-stylization method that standardizes feature maps of images to help produce consistent predictions. Image classification experiments on two benchmarks demonstrated that the proposed method greatly improves performance in both single-source and multi-source settings. We also show that the proposed method works effectively in person-reID experiments. In all experiments, our method significantly improves the performance.

1 INTRODUCTION

Deep neural networks (DNNs) has brought remarkable advances in a number of research areas such as visual recognition (Krizhevsky et al., 2012), image synthesis (Goodfellow et al., 2014), and reinforcement learning (Mnih et al., 2013). Most of successful models assume that training and test data are sampled under independent and identically distributed (i.i.d.) condition, which often does not hold in real-world environments unfortunately; a large error occurs when out-of-distribution data is given due to the distribution shift problem. To alleviate this problem, domain adaptation has been studied for learning domain-invariant models using fully labeled source data and target data with few or no labels. In many real applications, however, target domains are latent and data of the domains are not accessible accordingly. Domain generalization addresses this issue by learning models that well generalize to unseen domains, and has attracted increasing attention.

The mainstream research of domain generalization follows the flow of domain adaptation, and explores ways of aligning the distribution between features of multiple domains by adversarial training (Ghifary et al., 2015; Li et al., 2018a;b), reducing the maximum mean discrepancy (Muandet et al., 2013; Ghifary et al., 2016), or contrastive learning (Kim et al., 2021a). In domain adaptation, since the images of the target domain are available, it is obvious to reduce the target error by aligning the domain, but this cannot be guaranteed in domain generalization. In addition, it is unclear whether the target data will be mapped to the aligned features, and there is a high risk of overfitting the classifier to the source domains. Meanwhile, meta-learning frameworks recently have been proposed to increase the generalization ability through episodic training that separates source domains and simulates situations where the distribution shift occurs. Unfortunately, it is still difficult to completely avoid the same issue since these methods can be seen as extensions of distribution alignment methods.

In this work, we aim to address this overfitting issue and propose a method for learning a relaxed classifier rather than fitting a classifier to completely classify the class on the given source domains. Inspired by knowledge distillation, but revisiting it for domain generalization, we propose a regularization method that allows a model meaningfully wrong predictions that may occur in multiple domains. Specifically, we propose to exploit the ensemble of predictive distributions whose input

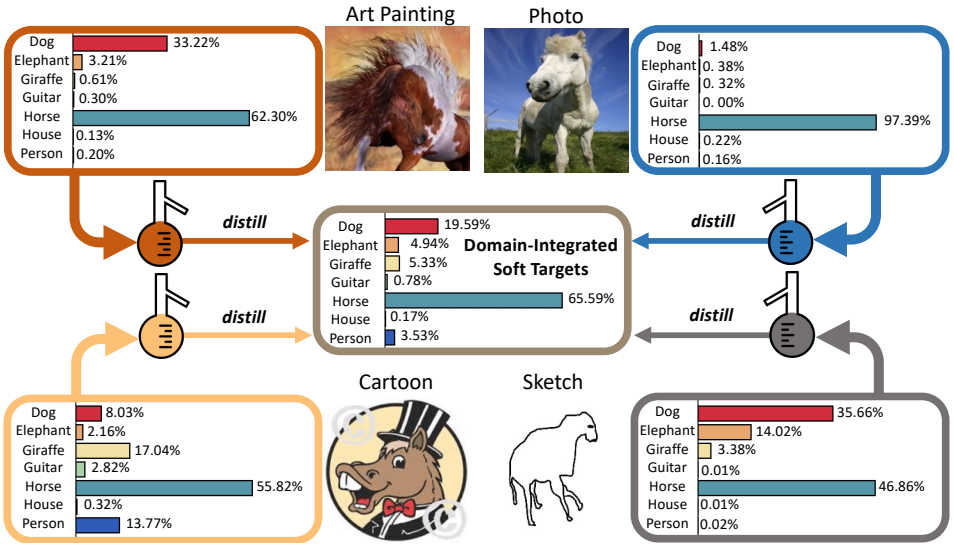

Figure 1: Illustration of cross domain ensemble distillation (XDED). To alleviate overfitting on source domains, XDED produces the soft target as a class-wise ensemble of predictive distributions of samples from different domains, which contains meaningful errors from multiple domains.

data have the same class label but belong to different domains as the knowledge and match it with each predictive distribution. We name our method cross domain ensemble distillation (XDED) and illustrate its main idea in Fig. 1. We remark that, unlike conventional knowledge distillation, XDED does not require multiple pretrained models, and employs a self-distillation manner that distills predictions obtained from a single network to itself. As a result, the proposed method increases the entropy of model predictions by penalizing the prediction of a sample with the ensemble which contains meaningful errors accumulated from multiple domains, which encourages the model to converge to wide local minima (Zhang et al., 2018; Cha et al., 2021b). Moreover, according to the theorem 1 proven by Cha et al. (2021a), wide minima leads to a small domain generalization gap, which explains that the proposed method can generalize well to unseen domains. We empirically show that our method contributes converging to a wide minima and improves generalization capability on unseen domains.

Since XDED is limited to regularization of the model only on source domains, there is still large room to further reduce the domain gap with the target domain. To this end, we also introduce an destylization technique well-suited for domain generalization, called UniStyle. UniStyle suppresses domain-specific style bias simply by standardizing intermediate feature maps of the image during both training time and *testing time*. Thanks to UniStyle, the model is able to produce style-consistent predictions in not only the source domains but also the target domain, which in turn greatly reduces the domain gap and boosts the effect of XDED.

We first demonstrate the effectiveness of our method on PACS (Li et al., 2017), a standard public benchmark for domain generalization. Our method significantly enhances the generalization ability of a model in both multi-source and single-source settings. We also validate the universality of the proposed method in various domain generalization scenarios by showing the improvements of image classification performance on the large-scale benchmark called DomainBed (Gulrajani & Lopez-Paz, 2021) and image retrieval performance for person-reID (Zheng et al., 2015; 2017). In all domain generalization experiments, our method achieves significant performance improvements.

2 Related Work

Domain Generalization. The goal of domain generalization is to learn domain-invariant features that are well-generalizable to the unseen target domain. Previous approaches suggested ways of matching distributions between different domains by adversarial feature alignment (Li et al.,

2018a;b) or reducing Maximum Mean Discrepancy (Muandet et al., 2013; Ghifary et al., 2016). Recently, meta-learning frameworks (Balaji et al., 2018; Li et al., 2019; Dou et al., 2019) have been investigated, and they simulate the domain shift by dividing the meta-train and meta-test domains from the original source domains. On the other hand, data augmentation methods have been proposed with the purpose of generating more diverse images beyond images of given source domains. For instance, CrossGrad (Shankar et al., 2018) perturbs images according to adversarial gradients induced by a domain classifier. L2A-OT (Zhou et al., 2020) learns a generator to map source data to synthetic domains by maximizing a divergence measure. FACT (Xu et al., 2021) mixes the amplitude spectrums of two images from a Fourier-based perspective.

Knowledge distillation and ensemble. Knowledge distillation, which is mainly devised for model compression, aims to transfer the knowledge of a deep model to a shallow model. As a seminal example, Hinton et al. (2015) encourages the student model to imitate class logits of the teacher model, which contain richer information than one-hot labels. While a myriad of studies have been investigated for various purposes such as cross-modality learning (Tian et al., 2020) and metric learning (Park et al., 2019; Kim et al., 2021b) and network regularization (Xu & Liu, 2019; Zhang et al., 2019; Yun et al., 2020). Unlike the conventional teacher-student framework, these network regularization methods called self-distillation distill own knowledge from their model itself and enforce consistency regularization between the original data and other data. Meanwhile, methods applying knowledge distillation (Meng et al., 2018; Zhou et al., 2021; Feng et al., 2021) have been proposed for domain adaptation, the task most closely related to domain generalization. Following the teacher-student training scheme of conventional knowledge distillation, they train several teacher models in the source domains and ensemble them to distill to student model. However, these approaches require large resources and training time since they require multiple pretrained teacher model. In addition, they are difficult to extend to domain generalization because they utilize target images along the strategy of domain adaptation. Our method simply yet effectively improves the generalization capability of the model without the need for target images and several teachers.

Bias towards styles. As recent studies (Geirhos et al., 2019; Brendel & Bethge, 2019) demonstrate that deep neural networks overly depend on a strong bias towards styles, it is also confirmed that the visual domain is closely related to its own style in domain generalization community (Zhou et al., 2021). Therefore, previous approaches have been proposed to define styles as bias and attempt to remove the dependency through augmentation (Zhou et al., 2021) or adversarial training (Nam et al., 2021). Distinct from them, in this paper, we propose a simple but effective de-stylization technique for domain generalization.

3 Our Approach

The goal of domain generalization is to learn domain-invariant representations from multiple source domains to generalize to unseen target domains. To achieve this goal, approaches to align the distributions of source domains and train discriminative classifier have been mainly explored. However, no information about the target domain is given in the domain generalization setting, so they have a high risk of overfitting to source domains.

In this work, we propose a knowledge distillation method for domain generalization that regularizes the model to mitigate this issue by learning meaningful wrong predictions accumulated from multiple domains. We also introduce an image de-stylization technique that maximizes the effect of our distillation and helps to produce consistent predictions not only in the source domain but also in the target domain. Lastly, we provide theoretical interpretation on how each component of our method leads to a small domain gap with empirical evidences.

3.1 Cross Domain Ensemble Distillation

Review of knowledge distillation. The goal of knowledge distillation (KD) (Hinton et al., 2015) is to transfer knowledge from a teacher model t to a student model s, usually a wide and deep model to a smaller one, for the purposes of model compression or model regularization. Given input data x and its label $y \in \{1, \cdots, C\}$, we denote the output logit of model as $z(x) = [z_1(x), \cdots, z_C(x)]$.

The posterior predictive distribution of data x is then formulated as:

$$P(y|x;\theta,\tau) = \frac{\exp(z_y(x)/\tau)}{\sum_{i=1}^{C}\exp(z_i(x)/\tau)}, \tag{1}$$

where the model is parameterized by θ and τ is a temperature scaling parameter. Knowledge distillation enforce to match the predictive distributions of s and t. Specifically, it is achieved by minimizing the Kullback-Leibler (KL) divergence between their predictive distributions as follows:

$$\mathcal{L}_{\text{KD}}(X;\theta_s) = \sum_{x_i\in X}\sum_{c=1}^{C} D_{KL}(P(c|x_i;\theta_t,\tau)||P(c|x_i;\theta_s,\tau)), \tag{2}$$

where X is a batch of input data, θ_t and θ_s are the parameters of a teacher and a student, respectively.

Cross domain ensemble distillation. We propose a new knowledge distillation method for domain generalization called cross domain ensemble distillation (XDED), which aims to transfer complementary knowledge using ensembles of logits from different domains. Unlike conventional KD, our method does not require an additional network which increases training complexity (*e.g.*, extra parameters and training time) but distills the ensemble knowledge constructed by multiple samples to the model itself in a form of self-knowledge distillation. More specifically, XDED produces the ensemble by averaging multiple logits whose class labels are the same in a mini-batch. The driving rationale behind the ensemble is to encode more complementary knowledge since different domains manifest different inter-class relations (*e.g.*, as shown in Fig. 1, the predictive distribution of Cartoon has high probability on class Person, but that of Sketch has high probability on class Dog.). As a result, each sample not only contributes to constructing the complementary knowledge by providing its own or its domain-specific information but also is supervised with that knowledge to learn domain-invariant information. Formally, let X_y denote the set of samples that have the same class label y in a mini-batch. Then, we obtain an ensemble of logits from X_y by simply taking an average as:

$$\bar{z}(X_y) = \sum_{x_i\in X_y}\frac{z_{x_i}}{|X_y|}. \tag{3}$$

Then, the predictive distribution from X_y can be defined as:

$$\bar{P}(c|X_y;\theta,\tau) = \frac{\exp(\bar{z}_c(X_y)/\tau)}{\sum_{i=1}^{C}\exp(\bar{z}_i(X_y)/\tau)}, \tag{4}$$

Therefore, the loss function of XDED is defined as follows:

$$\mathcal{L}_{\text{XDED}}(X_y;\theta) = \sum_{x_i\in X_y}\sum_{c=1}^{C} D_{KL}(\bar{P}(c|X_y;\hat{\theta},\tau)||P(c|x_i;\theta,\tau)), \tag{5}$$

where $\hat{\theta}$ is a fixed copy of the parameter θ. Following Miyato et al. (2018), we stop the gradient to be propagated through $\hat{\theta}$ to prevent the model from falling into some trivial solutions. To sum up, we set our objective function as

$$\min_{\theta} L_\theta = \mathcal{L}_{\text{CE}}(X,Y;\theta) + \lambda\sum_{c=1}^{C}\mathcal{L}_{\text{XDED}}(X_c;\theta), \tag{6}$$

where X is a batch of input images, Y is a batch of corresponding class labels, $\mathcal{L}_{\text{CE}}$ denotes the vanilla cross-entropy loss, and λ is a hyperparameter to balance $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{XDED}}$. Unless specified otherwise, λ and τ are 5.0 and 4.0 throughout this paper.

3.2 UniStyle: Removing and Unifying Style Bias

For regularizing the model to produce style-consistent predictions, we propose a de-stylization technique which is well-suited for domain generalization. Since domain-specific styles are not expected to be held at test time, we propose UniStyle to prevent the model from being biased towards the domain-specific styles, thus, reduce the domain gap with the target domain.

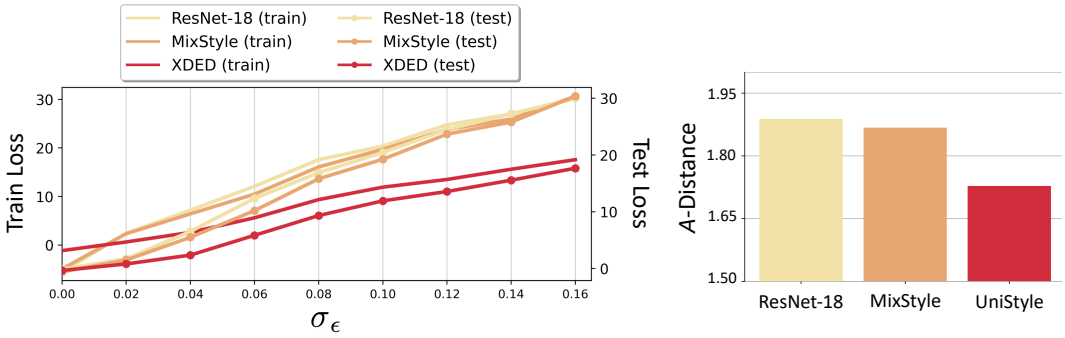

Figure 2: Analysis of our framework. **Left**: Train/Test losses versus the weight perturbation with varying σ_ϵ. Note that the loss values are log-scaled, **Right**: The divergence ($\mathcal{A}$-distance) between the source domains and the target domain.

More specifically, following recent approaches related to style transfer (Huang & Belongie, 2017; Nam & Kim, 2018; Lee et al., 2019), we first represent a neural style as statistics of intermediate featuremaps from the feature extractor. Formally, let $F \in \mathbb{R}^{C\times H\times W}$ denote an intermediate feature map of an image. Then, a neural style of the image is represented as the combination of channel-wise mean $\mu(F) \in \mathbb{R}^C$ and standardization $\sigma(F) \in \mathbb{R}^C$ of F as:

$$\mu_c(F) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{c,h,w}, \tag{7}$$

and

$$\sigma_c(F) = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(F_{c,h,w} - \mu_c(F))^2}, \tag{8}$$

where $\mu(F) = [\mu_1(F), \cdots, \mu_C(F)]$ and $\sigma(F) = [\sigma_1(F), \cdots, \sigma_C(F)]$. Next, we simply standardize each feature to have constant channel-wise statistics, μ_W and σ_W as:

$$\text{UniStyle}(F) = \sigma_W \frac{F - \mu(F)}{\sigma(F)} + \mu_W, \tag{9}$$

where we select $\mu_W = \mathbf{0}$ and $\sigma_W = \mathbf{1}$ (*i.e.*, zero-mean standardization). We observed that UniStyle is effective when being applied at multiple early layers, which is alinged with recent studies (Huang & Belongie, 2017; Dumoulin et al., 2017) suggesting that the style information is usually captured at the early layers. Further analysis on UniStyle is supplemented in the supplementary material A.1.

3.3 Theoretical Interpretation

In this section, we provide theoretical interpretation on how our framework leads to a smaller domain generalization gap, starting from the theorem related to domain adaptation (Ben-David et al., 2007; 2010). This theorem shows that the expected risk on the target domain is bounded by the expected risk on the source domain and the divergence between the target domain and the source domain.

Meanwhile, to find a model parameter $\theta \in \Theta$ for domain generalization, Cha et al. (2021a) recently introduced a robust empirical loss as:

$$\hat{\varepsilon}_S^\gamma(\theta) := \max_{||\Delta||\leq\gamma} \hat{\varepsilon}_S(\theta + \Delta) \tag{10}$$

where $\hat{\varepsilon}_S(\theta)$ is an empirical risk over source domains S and γ is a radius which defines neighbor parameters of θ. Then, Cha et al. (2021a) theoretically showed that finding wide local minima reduces the domain gap through the theorem below:

Theorem 1. *Consider a set of N covers* $\{\Theta_k\}_{k=1}^N$ *such that the hypothesis space* $\Theta \subset \cup_k^N \Theta_k$ *where* $diam(\Theta) := \sup_{\theta,\theta'\in\Theta} ||\theta - \theta'||_2, N := \lceil (diam(\Theta)/\gamma)^d \rceil$ *and* d *is dimension of* Θ. *Let* v_k *be a*

VC dimension of each Θ_k. Then, for any $\theta \in \Theta$, the following bound holds with probability at least $1-\delta$,

$$\varepsilon_T(\theta) < \hat{\varepsilon}_S^{\gamma}(\theta) + \frac{1}{2I}\sum_{i=1}^{I} \textit{Div}(S_i, T) + \max_{k\in[1,N]} \sqrt{\frac{v_k \ln(m/v_k) + \ln(N/\delta)}{m}}, \tag{11}$$

where $m = nI$ is the number of training samples and Div(S_i, T) is the divergence between the source domain S_i and the target domain T.

We remark that, in Eq. (11), the test loss $\varepsilon_T(\theta)$ is bounded by three terms: the robust empirical loss $\hat{\varepsilon_S^{\gamma}}(\theta)$ and the divergence $\text{Div}(S_i, T)$. In the rest of this section, we show that our framework lowers both $\hat{\varepsilon_S^{\gamma}}(\theta)$ and $\text{Div}(S_i, T)$ with the empirical evidences. Following empirical evidences are produced from the PACS dataset by training models on three source domains (*i.e.*, "Cartoon", "Sketch", and "Photo") and evaluating them on a target domain (*i.e.*, "Art Painting").

XDED lowers the robust empirical loss. To demonstrate that XDED promotes wide local minima, we quantify how wide each model converges to a local minima by measuring changes of loss value between θ and its neighborhoods, assuming that promoting wide local minima would have smaller changes. More specifically, following Zhang et al. (2018); Cha et al. (2021b), we measure the training losses of the learned models before and after adding Gaussian noise to model parameters while varying the standard deviation of the noise σ_ϵ (*i.e.*, $\mathcal{L}_{\text{CE}}(X, Y; \theta+\epsilon)$ where $\epsilon \sim N(0, \sigma_\epsilon)$). As shown in Figure. 2 (left), the results show that XDED demonstrates its robustness against the weight perturbation with smaller loss changes. With the benefit from entropy regularization approaches (Pereyra et al., 2017; Szegedy et al., 2016; Zhang et al., 2018; Cha et al., 2021b) to finding wide local minima, XDED promotes wide local minima by penalizing mismatches of predictive distributions between samples from different domains in the context of domain generalization. To sum up, we empirically show that XDED contributes to smaller domain generalization gap by reducing the robust empirical loss.

UniStyle lowers the domain discrepancy. To examine the effectiveness of the proposed whitening in reducing the divergence $\text{Div}(S_i, T)$, we adopt $\mathcal{A}$-distance (Kifer et al., 2004; Ben-David et al., 2010) as a measure. However, since computing the exact $\mathcal{A}$-distance is generally intractable, following Long et al. (2015), we calculate an approximated version of $\mathcal{A}$-distance between features from the target domain and source domains, which is defined as $\hat{d}_{\mathcal{A}} = 2(1 - 2\epsilon_{\text{svm}})$ where ϵ_{svm} is the generalization error of a SVM-based two-class classifier trained to distinguish the domain membership of input features. As shown in Figure. 2 (right), we observe that our whitening clearly lowers the distance with negligible computational overheads when compared to the vanilla ResNet-18 and MixStyle (Zhou et al., 2021).

In conclusion, we empirically show that the proposed methods, XDED and UniSyle, lead to a smaller domain generalization gap by reducing robust empirical loss and domain discrepancy, respectively. Their effects are complementary, and the combination of our two simple methods can significantly improve the domain generalization ability.

4 Experiments

4.1 Generalization in image classification

In this section, to demonstrate the superiority of our framework, we evaluate the proposed framework on the task of domain generalization in image classification.

4.1.1 Conventional Setting

Experimental setup. Specifically, for fair comparison, we follow the leave-one-domain-out protocol (Li et al., 2017) where we train a model on three domains and evaluate it on the remaining domain for multi-source domain generalization. For single-source domain generalization, we train a model on single domain and evaluate it on the other three domains. We use ResNet-18 (He et al., 2016) as backbone of our model, and our UniStyle technique is applied to output feature maps of the first and second residual blocks. For the benchmark dataset, we employ the PACS (Li et al., 2017)

Table 1: Leave-one-domain-out generalization results on PACS

Methods	Art	Cartoon	Photo	Sketch	Average
DeepAll	77.0	75.9	96.0	69.2	79.5
MMD-AE (Li et al., 2018a)	75.2	72.7	96.0	64.2	77.0
CCSA (Motiian et al., 2017)	80.5	76.9	93.6	66.8	79.4
JiGen (Carlucci et al., 2019)	79.4	75.3	96.0	71.6	80.5
CrossGrad (Shankar et al., 2018)	79.8	76.8	96.0	70.2	80.7
MASF (Dou et al., 2019)	80.2	77.1	94.9	71.6	81.0
Epi-FCR (Li et al., 2019)	82.1	77.0	93.9	73.0	81.5
MetaReg (Balaji et al., 2018)	83.7	77.2	95.5	70.3	81.7
EISNet (Wang et al., 2020)	81.8	76.4	95.9	74.3	82.1
L2A-OT (Zhou et al., 2020)	83.3	78.2	96.2	73.6	82.8
SagNet (Nam et al., 2021)	83.5	77.6	95.4	76.3	83.2
SelfReg (Kim et al., 2021a)	82.3	78.4	96.2	77.4	83.6
MixStyle (Zhou et al., 2021)	84.1	78.8	96.1	75.9	83.7
L2D (Wang et al., 2021)	81.4	79.5	95.5	80.5	84.2
FACT (Xu et al., 2021)	85.3	78.3	95.1	79.1	84.5
DSON (Seo et al., 2020)	84.6	77.6	95.8	**82.2**	85.1
RSC (Huang et al., 2020)	83.4	80.3	95.9	80.8	85.1
Ours	**85.6**	**84.2**	**96.5**	79.1	**86.4**

Table 2: Single-source domain generalization accuracy (%) on PACS with a ResNet-18 backbone. (A: Art Painting, C: Cartoon, S:Sketch, P:Photo).

Methods	A→C	A→S	A→P	C→A	C→S	C→P	S→A	S→C	S→P	P→A	P→C	P→S	Avg
ResNet-18	62.3	49.0	95.2	65.7	60.7	83.6	28.0	54.5	35.6	64.1	23.6	29.1	54.3
JiGen (Carlucci et al., 2019)	57.0	50.0	96.1	65.3	65.9	85.5	26.6	41.1	42.8	62.4	27.2	35.5	54.6
MixStyle (Zhou et al., 2021)	65.5	49.8	96.7	69.9	64.5	85.3	27.1	50.9	32.6	67.7	38.9	39.1	57.4
RSC (Huang et al., 2020)	62.5	53.1	96.2	68.9	**70.3**	85.8	37.9	56.3	47.4	66.3	26.4	32.0	58.6
SelfReg (Kim et al., 2021a)	65.2	55.9	96.6	72.0	70.0	87.5	37.1	54.0	46.0	67.7	28.9	33.7	59.5
SagNet (Nam et al., 2021)	67.1	56.8	95.7	72.1	69.2	85.7	41.1	62.9	46.2	69.8	35.1	40.7	61.9
Ours	**74.6**	**58.1**	**96.8**	**74.4**	69.6	**87.6**	**43.3**	**65.6**	**50.3**	**71.4**	**54.3**	**51.5**	**66.5**

that is a widely-used benchmark for domain generalization in image classification. PACS consists of 9,991 images over 4 domains: Art Painting, Cartoon, Photo and Sketch.

Results of multi-source domain generalization. As summarized in Table. 1, our method clearly outperforms the latest methods which are dedicated to domain generalization. Except for the case of Sketch domain, our method achieves the best accuracies among other competing methods. Especially, for Cartoon domain, our method exceeds RSC (Huang et al., 2020), the second best method, by about 4.0%. The most challenging domain for our method is Sketch. We conjecture this is because images of Sketch are colorless and our UniStyle removes too much information at inference time. Nevertheless, we remark that not only Sketch is commonly one of the most difficult domains for other methods too. In sketch, each method usually shows the lower performance than its average accuracy. Nevertheless, we remark that bour method shows its superiority over other methods on overall performance.

Results of single-source domain generalization. Thanks to the simple design of our proposed method which does not explicitly require domain labels, our framework can be transparently incorporated with single-source domain generalization where we only have access to a single source domain during training. Therefore, to further evaluate the impact of our framework on single-source domain generalization, our model is trained on each single domain of PACS and evaluated on the remaining target domains. As shown in Table. 2, our model significantly outperforms other baselines by 8.7% in average accuracy. Besides, in all cases except for the case of $C \rightarrow S$, our model shows its superiority in performance. Even though only a single domain is provided during training, we conjecture this interesting result is attributed to the existence of inherent variation between intra-domain samples, which is also aligned with the observation of multiple sub-domains in a single domain (Zhou et al., 2021), and the ability of our framework to exploit that fine-grained relations between samples.

Model selection: leave-one-domain-out cross-validation							
Algorithm	**CMNIST**	**RMNIST**	**VLCS**	**PACS**	**OfficeHome**	**TerraIncognita**	**Avg**
ERM	36.7	97.7	77.2	83.0	65.7	41.4	66.9
IRM	40.3	97.0	76.3	81.5	64.3	41.2	66.7
GroupDRO	36.8	97.6	77.9	83.5	65.2	44.9	66.7
Mixup	33.4	97.8	77.7	83.2	67.0	**48.7**	67.9
MLDG	36.7	97.6	77.2	82.9	66.1	46.2	67.7
CORAL	39.7	97.8	**78.7**	82.6	**68.5**	46.3	**68.9**
MMD	36.8	97.8	77.3	83.2	60.2	46.5	66.9
DANN	40.7	97.6	76.9	81.0	64.9	44.4	67.5
CDANN	39.1	97.5	77.5	78.8	64.3	39.9	66.1
MTL	35.0	97.8	76.6	83.7	65.7	44.9	67.2
SagNet	36.5	94.0	77.5	82.3	67.6	47.2	67.5
ARM	36.8	**98.1**	76.6	81.7	64.4	42.6	66.7
VREx	36.9	93.6	76.7	81.3	64.9	37.3	65.1
RSC	36.5	97.6	77.5	82.6	65.8	40.0	66.6
Ours	**46.5**	97.7	74.8	**83.8**	65.0	42.5	68.4

Table 3: Domain generalization accuracy (%) on DomainBed. The results compare fifteen methods including ours across six domain generalization benchamark datasets. Note that we adopt leave-one-domain-out cross-validation as a model selection criteria.

4.1.2 DOMAINBED

Experimental setup. We also conduct extensive experiments on the DomainBed (Gulrajani & Lopez-Paz, 2021) which is a testbed for domain generalization to compare state-of-the-art methods across several benchmark datasets. The rationale behind the DomainBed is that the domain generalization performances are too much dependent on the hyperparameter tuning. Therefore, for a fair comparison, we follow the its rigorous protocols for training and evaluation.

Results. As shown in Table. 3, our method generally shows better or competitive performances and ranks second out of 15 methods on average. Especially, on CMNIST, our method substantially outperforms other competing methods. We conjecture this performance boost is attributed to the de-stylization of UniStyle since CMNIST is designed to simulate the domain shift via background colors which are highly correlated with visual styles rather than other factors such as shape.

4.2 GENERALIZATION IN PERSON RE-ID

In this section, we further evaluate our framework on the person re-identification (re-ID), which is the task of matching pedestrians across non-overlapping camera views. Considering each camera as a source domain, learning invariance of each identity across different domains is the key to success in person re-ID.

Experimental setup. Here, we address domain generalization for person re-ID, where the test data is collected from cameras of the unseen dataset rather than from those of the training dataset. Specifically, the model trained to match people in the source dataset is then evaluated by how well it matches pedestrian data of the unseen test set, which are disjoint from those of the source dataset. For datasets, we adopt two widely-used benchmarks: Market1501 (Zheng et al., 2015) and DukeMTMC-reID (Duke) (Ristani et al., 2016; Zheng et al., 2017). We use 32,668 images of 1,501 identities collected from 6 cameras and 36,411 images of 1,812 identities from 8 cameras for Market1501 and Duke, respectively. As performance measures, we adopt mean average precision (mAP) and Recall@K (R@K). Following the prior work (Zhou et al., 2021), we adopt ResNet-50 (He et al., 2016) as a backbone architecture. In these experiments, we apply UniStyle to the 1st, 2nd and 3rd residual blocks of a model.

Comparison to other regularization methods. As shown in Table. 4, our framework substantially outperforms other methods in mAP and Recall@1. Although RandomErase and Dropblock are known to be effective for learning discriminative features, they both fail to improve performance when encountered unseen domain data. Furthermore, by exploiting inter-class relations provided by

Methods	Market1501→Duke				Duke→Market1501			
	mAP	**R@1**	**R@5**	**R@10**	**mAP**	**R@1**	**R@5**	**R@10**
ResNet-50	19.3	35.4	50.3	56.4	20.4	45.2	63.6	70.9
RandomErase (Zhong et al., 2020)	14.3	27.8	42.6	49.1	16.1	38.5	56.8	64.5
DropBlock (Ghiasi et al., 2018)	18.2	33.2	49.1	56.3	19.7	45.3	62.1	69.1
MixStyle (Zhou et al., 2021)	23.4	43.3	**58.9**	**64.7**	24.7	53.0	**70.9**	**77.8**
Ours	**27.4**	**49.3**	56.0	59.5	**30.1**	**59.0**	67.0	71.5

Table 4: Generalization results on the cross-dataset person re-ID task.

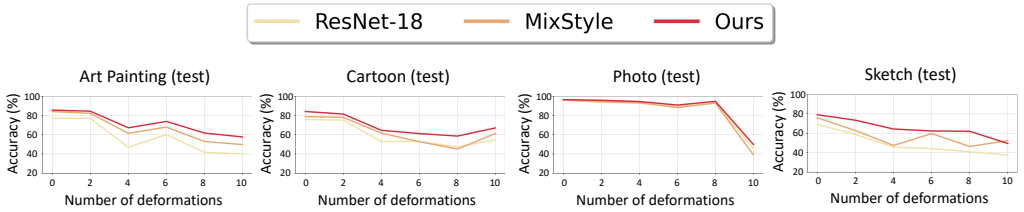

Figure 3: Test accuracy (%) on target domains with input deformations. All models are trained under the multi-source domain generalization setting.

method	Accuracy (%)
Label smoothing	79.9
MixUp	78.5
Ours	86.4

Table 5: Comparison with other regularization methods exploiting soft targets.

different cameras, our framework shows its superiority over MixStyle which is designed for domain generalization but utilizes one-hot labels only, resulting in ignoring inter-class relations.

4.3 In-Depth Analysis

Generalization on target domains with input deformations. To further evaluate the generalization performance of the proposed method on more challenging, we simulate gradually increasing domain shift by adding deformations on images of the target domain. Specifically, we transformed the image in target domain by applying multiple augmentations defined in RandAugment (Cubuk et al., 2020), and gradually increased the number of them. For a fair comparison, we use same augmentation operations for each setting. As shown in Fig. 3, our method demonstrates its better generalization ability even in more challenging conditions with large domain shift, which is attributed to both promoting wide local minima and reducing the domain gap by reducing style bias. This result also supports

How much does soft target matter by itself? Considering our usage of soft targets, we investigate whether the perfomance boosts are merely due to the usage of soft targets. Thus, when compared to Label smoothing and Mixup which both exploit soft targets in their own ways, Table. 5 shows our method substantially outperforms them. It is because their soft targets are not able to capture the domain knowledge.

5 Conclusion

In this paper, we presented a simple yet effective method for domain generalization. Distinct from existing techniques which risks of overfitting on source domains, XDED allows meaningful errors of a model in a form of knowledge distillation, helping the model promote wide local minima. Besides, the proposed UniStyle suppresses domain-specific style so that it helps the model produce style-consistent predictions, resulting in a reduced domain gap. Furthermore, we also provide theoretical interpretation on how each component contributes to improved generalization ability with empirical evidences. Through the extensive experiments, our framework achieves significant performance improvements.

REFERENCES

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 2010.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2021a.

Sungmin Cha, Hsiang Hsu, Taebaek Hwang, Flavio P Calmon, and Taesup Moon. Cpr: Classifier-projection regularization for continual learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2021b.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Hao-Zhe Feng, Zhaoyang You, Minghao Chen, Tianye Zhang, Minfeng Zhu, Fei Wu, Chao Wu, and Wei Chen. Kd3a: Unsupervised multi-source decentralized domain adaptation via knowledge distillation. In *Proc. International Conference on Machine Learning (ICML)*, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2018.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *In Very Large Databases (VLDB)*, 2004.

Daehee Kim, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021a.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Embedding transfer with label relaxation for improved metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2012.

HyunJae Lee, Hyo-Eun Kim, and Hyeonseob Nam. Srm: A style-based recalibration module for convolutional neural networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018a.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2018b.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proc. International Conference on Machine Learning (ICML)*, 2015.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang. Adversarial teacher-student learning for unsupervised domain adaptation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NeurIPS Deep Learning Workshop*, 2013.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proc. International Conference on Machine Learning (ICML)*, 2013.

Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2018.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR Workshop*, 2017.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing (TIP)*, 2021.

A APPENDIX

This supplementary material presents additional ablation studies on the proposed framework and its implementation details, all of which are omitted from the main paper due to the space limit. A.1 presents additional analysis results on the proposed method. A.2 provides the implementation details of the proposed method in each task.

A.1 ADDITIONAL ABLATION STUDIES

Learning Acceleration. Our framework enables faster convergence of the learning process. In generalization on PACS, our effectiveness of learning acceleration is demonstrated. As shown in Fig. 4, our framework reaches higher performance with less iterations compared to both vanilla method and MixStyle (Zhou et al., 2021). Although the vanilla method requires many iterations for convergence to cover the domain gaps between source domains, our framework accelerates the learning process via encouraging the model to consider different inter-class relations among source domains. Since MixStyle (Zhou et al., 2021), as an augmentation method, aims to synthesize novel styles via mixing statistics at the feature level, it is inherently limited to require many iterations to get many augmented styles for its best performance.

Where to apply UniStyle? We remark that all the quantitative results above are produced with the ResNet (He et al., 2016) as a backbone network and the proposed UniStyle can be applied after arbitrary intermediate layers of the backbone as a plug-and-play module. Therefore, to investigate the impact of where UniStyle is applied, we evaluate the generalization performance in image classification and person re-ID while varying the locations of where the operation is applied. For a baseline, MixStyle (Zhou et al., 2021) is adopted and compared with our UniStyle since they share the commonality of being applied to multiple intermediate featuremaps. For brevity, let RES# denote the indexes of residual blocks where the specified operation is applied (*e.g.*, RES12 means the operation is applied after both the first and second residual blocks). As shown in Table. 6, we observe that UniStyle and MixStyle have a similar trend in both tasks. First, since early layers are known to capture low-level features such as texture or edge information, it is pertinent for both UniStyle and MixStyle to be applied after early layers to remove style bias and synthesize novel styles, respectively. However, UniStyle achieves the best performances on both tasks, which indicates the importance of removing and matching style bias rather than augmenting novel styles. Next, on the contrary, both operations lead critical performance drop when incorporated with the last layer, RES4, since late layers are known to address semantic information (*i.e.*, their statistics would be highly correlated with target labels). In detail, MixStyle perturbs the statistics by interpolating the those of two different instances that may have different labels, whereas UniStyle normalizes the featuremap of all samples regardless of their labels, resulting in more critical performance drop. Note that only UniStyle is applied and the cross domain ensemble distillation is excluded in results of Table. 6.

Universality of our framework. We remark that the proposed framework can be incorporated with any other methods thanks to its simplicity without requiring any additional modules. Therefore, to demonstrate its universality, we evaluate generalization accuracy of baseline methods incorporated with our framework on PACS. As shown in Table. 8, when baselines are incorporated with our framework, the performances are consistently improved by 3.4%p on average. Note that no hyperparameter search was conducted to find the best combination of our framework and the baseline methods.

Ablation study on the effect of the proposed components. We conduct ablation study to examine the effect of the proposed components. Therefore, we evaluate the generalization performances in image classification and person re-ID tasks. As shown in Table. 7, the proposed component consistently boosts the generalization performances in both tasks.

A.2 IMPLEMENTATION DETAILS

For the task of domain generalization in image classification, we train the models using the sgd optimizer with the cosine learning decay (Loshchilov & Hutter, 2016) and initial learning rate of 10^{-3}. They are learned for 100 epochs. For batch construction, we use the batch size of 64 and

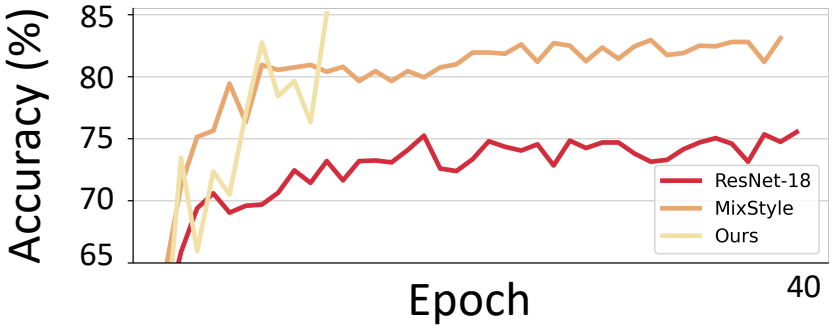

Figure 4: The learning curve on the target domain. On PACS, models are trained on the source domains (Cartoon, Sketch and Photo) and evaluated on the target domain (Art Painting).

Table 6: Ablation study on where to apply UniStyle in the ResNet architecture.

(a) Image classification on PACS

Model	Accuracy (%)
ResNet-18	79.5
+ MixStyle (RES1)	80.1
+ UniStyle (RES1)	81.5
+ MixStyle (RES12)	81.6
+ UniStyle (RES12)	**82.9**
+ MixStyle (RES123)	82.8
+ UniStyle (RES123)	82.4
+ MixStyle (RES1234)	75.6
+ UniStyle (RES1234)	12.8

(b) Person re-ID from Market1501→Duke

Model	mAP (%)
ResNet-50	19.3
+ MixStyle (RES1)	22.6
+ UniStyle (RES1)	22.8
+ MixStyle (RES12)	23.8
+ UniStyle (RES12)	22.8
+ MixStyle (RES123)	22.0
+ UniStyle (RES123)	**24.0**
+ MixStyle (RES1234)	10.2
+ UniStyle (RES1234)	0.2

Table 7: Ablation study on the proposed components.

(a) Image classification on PACS

Model	Accuracy (%)
ResNet-18	79.5
+ UniStyle	82.9
+ UniStyle + XDED	86.4

(b) Person re-ID from Market1501→Duke

Model	mAP (%)
ResNet-50	19.3
+ UniStyle	24.0
+ UniStyle & XDED	27.4

Methods	**Art**	**Cartoon**	**Photo**	**Sketch**	**Average**
JiGen (Carlucci et al., 2019)	79.4	75.3	96.0	71.6	80.5
Ours + JiGen	85.3	79.2	95.9	79.2	84.9
RSC (Huang et al., 2020) (*our imple*)	82.8	77.6	95.7	78.6	83.7
Ours+RSC	84.7	81.3	96.2	82.2	86.1

Table 8: Universality of our proposed framework. Generalization accuracy (%) on PACS

sample 16 instances per class for the proposed XDED. For the task of person re-ID, we also train the models using the sgd optimizer with initial learning rate of 0.05.